KEEP THIS COPY FOR REPRODUCTION PURPOSES

AD-A246 231

ATION PAGE

Form Approved
OMB No 0704-0188

average 1 hour per response, including the time for reviewing instructions, searching existing data sources ing the collection of information. Send comments regarding this burden estimate or any other aspect of this to Washington Headquarters Services, Directorate for information Operations and Reports, 1215 Jefferson of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
		Technical 1 May 88 - 31 May 1991

4. TITLE AND SUBTITLE

Scrawl Strips, and Letter or B-Letter Strips: Depicting Marginals of Scatter Plots

5. FUNDING NUMBERS

DAAG03-88-K-0045

6. AUTHOR(S)

John W. Tukey, and James G. Veitch

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Princeton University
Fine Hall
Washington Road
Princeton, NJ 08544-1000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

26098-MA-SDI

11. SUPPLEMENTARY NOTES

The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Data from a bivariate distribution is often graphically presented by using a scatter plot. Adding a suitable depiction of the marginal data distributions to the edges of the scatter plot allows interesting features of the marginal data distributions to be seen alongside the original bivariate data. We propose providing these marginal depictions by a modification of Quantile-Quantile plots (QQ plots) we call *scrawl strips*. We also propose adding a strip or axis showing location of the letter values, or the broadened letter values, which we call a *letter strip*, or *b-letter strip*, respectively. These two strips can be combined.

Since scrawl strips are modified QQ plots, they inherit useful features of QQ plots. These features include information about the shape of the marginal distribution, the presence and type of skewness, the presence of heavy tails, gaps and ties in the ordered data, appearances of bimodality or high shoulders, and assumptions of normality. This type of information is often either absent or depicted poorly, both in univariate competitors to the QQ plot and in the basic scatter plot.

14. SUBJECT TERMS

KEYWORDS: B-letter strip, Data analysis, Graphics, Letter strip, Quantile-quantile plot, Scatterplot, Scrawl strip.

15. NUMBER OF PAGES

23

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Scrawl Strips, and Letter or B-Letter Strips: Depicting Marginals of Scatter Plots

John W. Tukey, and James G. Veitch

Technical Report No. 299
Princeton University
Fine Hall
Washington Road
Princeton, NJ 08544-1000
and
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07929

ABSTRACT

Data from a bivariate distribution is often graphically presented by using a scatter plot. Adding a suitable depiction of the marginal data distributions to the edges of the scatter plot allows interesting features of the marginal data distributions to be seen alongside the original bivariate data. We propose providing these marginal depictions by a modification of Quantile-Quantile plots (QQ plots) we call *scrawl strips*. We also propose adding a strip or axis showing location of the letter values, or the broadened letter values, which we call a *letter strip*, or *b-letter strip*, respectively. These two strips can be combined.

Since scrawl strips are modified QQ plots, they inherit useful features of QQ plots. These features include information about the shape of the marginal distribution, the presence and type of skewness, the presence of heavy tails, gaps and ties in the ordered data, appearances of bimodality or high shoulders, and assumptions of normality. This type of information is often either absent or depicted poorly, both in univariate competitors to the QQ plot and in the basic scatter plot.

August 29, 1989

KEYWORDS: B-letter strip, Data analysis, Graphics, Letter strip, Quantile-quantile plot, Scatterplot, Scrawl strip.

This work was initially supported by U. S. Department of Energy contract DEAC0281ER10841 with Princeton University. This contract and a National Science Foundation grant MCS8304372 supported the computing equipment used at Princeton. J. W. Tukey's later work was supported by the Army Research Office (Durham), DAAG29-82-K-0178, DAAL03-86-K-0073, and DAAL03-88-K-0045.

Scrawl Strips, and Letter or B-Letter Strips: Depicting Marginals of Scatter Plots

John W. Tukey, and James G. Veitch

Technical Report No. 299
Princeton University
Fine Hall
Washington Road
Princeton, NJ 08544-1000
and
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07929

Introduction.

Data from a bivariate distribution is often graphically presented using a scatter plot. Normally, one is also interested in the univariate marginal distributions of the data; but, unfortunately, the scatter plot often does not show interesting features of these marginals. In particular, the shapes of the marginal distributions, ties in values in one variable, and gaps in the sequence of values of one variable may be difficult or impossible to see in the scatter plot. Bimodality in either marginal may or may not be easy to see. Many of these properties may be important, for example, the shapes of the marginal data distributions may have crucial implications in the type of data analysis we wish to select, especially when conventional assumptions of Gaussianity can be seen to be unjustified.

Ways to look at these marginal data distributions include plotting separate histograms (Tufte (1983)), rootograms, hanging or suspended rootograms (explained in Tukey (1965)), boxplots, displays of letter values (median, hinges, eighths, etc.), or Quantile-Quantile plots (QQ plots). The type of QQ plot we are interested in is a scatter plot of univariate ordered data (the order statistics or quantiles) against the expected (mean) or anticipated (median) values of these order statistics for a chosen distributional assumption (for example, Normal scores).

It would be desirable to present this univariate information alongside the scatter plot, so one could directly compare features of both the bivariate data distribution and the marginal data distributions. Other authors (see Chambers et al. (1983)) have used the "alongside" idea by including boxplots or jittered dot strips (discussed later) of the X data in a strip above the scatter plot, using the X-axis as a scale. Similarly, they plot the Y data alongside the vertical axis of the scatter plot. Tufte (1983) has advocated similar ideas.

Plausible reasons for depicting marginals.

Since presenting the marginal data distributions in strips alongside a scatter plot appears useful, it may be desirable to base the presentation of these marginals on a different method than, say, boxplots or histograms. Before choosing a method we should discuss what features are desirable in depicting a univariate distribution of data. Some desirable features in depicting univariate data include:

KEYWORDS: B-letter strip, Data analysis, Graphics, Letter strip, Quantile-quantile plot, Scatterplot, Scrawl strip.

This work was initially supported by U. S. Department of Energy contract DEAC0281ER10841 with Princeton University. This contract and a National Science Foundation grant MCS8304372 supported the computing equipment used at Princeton. J. W. Tukey's later work was supported by the Army Research Office (Durham), DAAG29-82-K-0178, DAAL03-86-K-0073, and DAAL03-88-K-0045.

(1) easy detection of outliers;

(2) easy semi-quantitative assessment of center and spread of the distribution;

(3) depiction of the general shape of the distribution (this includes presence and type of skewness or lack of symmetry, heavy or squashed tails, and possibly comparing the distribution with a reference distribution such as the Gaussian);

(4) detection of unusual gaps in the sequence of ordered data values;

(5) detection of ties in data values;

(6) noticeability of apparent bimodality, or of high shoulders compared to the reference distribution (which is appropriately called relative bimodality).

One wishes to choose a method which stresses those features of the marginal distributions the scatter plot will depict weakly or not at all. Outliers will often be clearer on a scatter plot than in the marginal distributions (for a discussion, see Tufte (1983), p.14), so item (1) seems to be the least important criterion in choosing a method of presentation. While one can try to visually estimate center and spread of the marginals from the scatter plot, one will usually get rather poor estimates, so item (2) is of moderate importance. By contrast, item (3) is almost never apparent on any scatter plot and is often of great interest, so it will be very important in choice of method of presentation. Of the other items, (4) and (5) will be difficult to see on the scatter plot, and item (6) may or may not be apparent. It seems to be generally true that techniques effective in depicting the broader aspects, such as items (2) or (3), do not serve so well in depicting narrower aspects, like items (4), (5), and (6), so we shall attack broader and narrower aspects separately.

Dealing with the broader aspects.

Histograms easily suggest center and spread of the distribution, albeit sometimes crudely. They seem often to give a reasonable indication of symmetry. The presence of heavier-than-Gaussian tails may be crucial to the type of analysis chosen (e.g. tests based on the usual correlation coefficient are likely to be non-conservative if the marginals are heavy-tailed), but histograms will not easily detect the presence of such heavy-tailedness. Histograms may give no information about gaps and ties, particularly if the interval sizes are not chosen cleverly, or if the gaps are small. In fact, presence of ties or gaps in the distribution may give rise to very different histograms, depending on the particular choice of interval widths and placement. An unlucky choice of interval widths or placement may also disguise the presence of bimodality (with regard to a rectangular distribution), or of high shoulders (with regard to the reference distribution). The heights of the bars are not strictly comparable, since they have different variances. Finally, it is not clear how to modify a histogram so as to fit in a small strip without losing its desirable characteristics. Rootograms retain many of these defects, though the heights of the rootogram bars are variance stabilized.

A modification of the suspended rootogram, see Tukey (1965), where one displays only the variation of the bars about the baseline, can be displayed in a narrow strip. This plot will do moderately well in presenting information about shape (item (2)), since the bars are now compared with an expected or fitted reference distribution, and the variance of the bar lengths is stabilized. Because only part of the suspended rootogram is displayed, such a plot requires some sophistication to understand and interpret effectively.

One can also present the marginal values along the edges of a scatter plot by marking or jittering these values (an explanation of jittering is in Chambers et al., Chapter 2 (1983)), or one may present information about these values by presenting boxplots . Both these alternatives will give some idea about symmetry, but will do worse than histograms in depicting other parts of item (3). Boxplots give a better feeling for estimating center and spread than merely marking or jittering the values. Marking or jittering will do better in detecting bimodality.

Much of the sample distribution shape can be summarized by the sequence of sample letter values, (cf. Tukey (1977), Chapter 2 of Hoaglin et al. (1983), and Chapter 10 of Hoaglin et al. (1985)). In our case, "M" denotes the sample median, the upper and lower hinges are denoted by "H", the upper and lower eighths are denoted by "E", the upper and lower sixteenths by "D", thirty-seconds by "C". We continue on in this fashion, where successive powers of two correspond to traversing the alphabet

backwards i.e. "B", "A", "Z", etc., until we hit the extremes of the sample. (Rules for dealing with fractions of an observation are set out in Tukey (1977), Hoaglin et al. (1983), or Hoaglin et al. (1985)). A simple presentation of the position of the letter values, including the median and hinges (shown in every boxplot), but also including the other letter values, tells us at least as much as the boxplot. Such information can be placed in a narrow strip or on an axis at the edge of the scatter plot, naturally called a *letter strip*. We use letter labels here (as in Tukey (1977), Hoaglin et al. (1983), and Hoaglin et al. (1985)) avoiding numerical labels naturally taken as fractions, for two reasons:

(a) *numerical fractions look like P-values, so that too many viewers would think about significance* rather than distribution shape, and

(b) no sequence decreasing by a constant ratio continues to have simple numbers over a wide range - thus using single letters will reduce clutter.

We have included a partial translation of the letters in Figure 1 for the convenience of readers unfamiliar with the use of letters. The letter values are estimates of the corresponding quantiles in the true (unknown) distribution, and can be made more stable - and, for many purposes, more useful - by using averages of the order statistics around a given letter value. We shall call such estimates *broadened letter values* and will display the broadened letter values by lower case letters, e.g. the broadened hinges will be denoted by "h". For a precise technical description of broadened letter values, see Mendoza (1984), and appendix. We shall naturally call the broadened letter values *b-letters*, and a strip containing b-letters a *b-letter strip*. (We have deferred to a referee here, but plan to use the word *bletter* elsewhere).

An example of a b-letter strip (note we have used lower case letters for b-letter values) is shown just beneath the top of the frame of Figure 1, where a QQ plot of 74 models of automobiles selected in model year 1979[1] has been made. For purposes of comparison, immediately below the b-letter strip is a letter strip, where the letters are ticked in using upper case. One can clearly see where the local gap at around 3000 pounds has shifted the median to the right, compared to the more stable broadened median. We easily see the left-hand e and h are farther from the broadened median m than their right-hand analogs, so there is some indication of left skewness in the shoulders of the distribution. When we look at the d's and c's (the left-hand c is coincident with the left-hand b), however, we see a clear indication of skewness to the right.

Figure 2 is a histogram of the same data, and shows that a histogram can often be considerably less helpful than a letter or b-letter strip, especially since it is insensitive to high shoulders alone. The circles in Figure 1 plot the b-letter values versus their anticipated means. We chose open circles so as to interfere only minimally with the data points plotted on the QQ plot. Because we feel a b-letter strip may be a useful addition to any QQ plot, we have replotted Figure 1 in Figure 6, where the distraction of the added letter strip has been removed.

Dealing with the other aspects.

QQ plots make steps towards remedying many of the remaining problems, but require more sophistication to interpret. Here, we shall consider only QQ plots of the ordered data against their Normal scores. However, knowledge about the data, or about assumptions one wishes to check, may sometimes make it appropriate to use order-statistic scores for some non-Gaussian distribution. Much of the information about distribution in a QQ plot lies in the curvilinear appearances shown by some plots. For example, antisymmetric appearing plots (using a symmetric reference) correspond to symmetric data distributions; if the reference is Gaussian, plots close to straight lines correspond to data which appears close to Gaussian, and by watching changes in the apparent slope of the plot, one can see whether tails of the distribution of the data are long or short tailed compared to the Gaussian, or whether there seem to be two or more areas of relative concentration. Since each data point appears on the plot, ties and gaps show up well. To gain such advantages, we use a modification of the QQ plot in our presentation of the marginals.

[1] This data came from the database distributed with the Bell Labs statistical package S and is tabulated in Chambers et al. (1983).

As an example, see Figure 4. Several things can be seen immediately from this plot. The data appears short-tailed with respect to a Gaussian reference on the low end, but the tail on the high end appears approximately Gaussian. Several large gaps are obvious. The biggest gaps are at 3000 and 3500 pounds, with other gaps noticeable at 2500 and 4000 pounds, and less obviously noticeable at 3800 pounds. Even if these gaps were removed by rigidly moving sections of the plot together horizontally, high shoulders with disproportionate concentrations near 2000 and 3400 pounds would still remain clear (from the two intervals of steeper slope) and bimodality would tend to be suspected. There is a tie between the weights of the second and third lightest model of car.

Scrawl strips.

As mentioned earlier, if one uses QQ plots to present the marginal distributions of bivariate data alongside the scatter plot, the natural idea is to put each QQ plot in a strip adjacent to the X and Y axes respectively, using the same scales for the data on both the strips and the scatter plot. The strip on the top (or bottom) thus plots X data along the X-axis scale, the strip on the left (or right) side thus plots Y data against the Y-axis scale.

This creates a problem. If one increases the size of the scatter plot relative to the QQ plots, on a piece of paper or display device of fixed size, one runs the risk of losing the visual information in the QQ plots due to a much reduced scale on the Q-axes corresponding to the expected order statistics, so the problem is: how does one reduce scale on the Q-axes while keeping the visual information in the QQ plot obvious? *Scrawl strips.* are one such method of presenting this information about the marginals. The idea is to preserve the visual information, but drastically reduce the physical space taken by the plot. The technique we focus on here is to first "fold" the original QQ plot, and then to present this new plot together with the scatter plot[2].

Here is a detailed description of how one might make a such a plot by hand:

(1) Prepare a the QQ plot by plotting the X data along the X-axis, and the expected order statistics of the chosen reference distribution on the other axis. As a example, see Figure 1.

(2) Choose an positive integer N. Divide the plot into N horizontal strips, each equally deep. Number these strips in ascending order, 1 to N, from lowest to highest on the plot. "Fold" the top strip, N, onto strip $(N-1)$ by reflecting all points in strip N in the dividing line between the two strips (see Figure 3).

(3) Repeat the process for the new top strip $(N-1)$ and strip $(N-2)$. Continue until only the bottom strip is left. The entire QQ plot is still present, but folded onto strip 1, with a consequent reduction in height. The final plot is called a *scrawl strip*. (We could call it a Vespal plot, as the common hornet *Vespa*, folds its wings at rest in a similar way to the folding of the QQ-plot). The reader may wish to compare the bottom strip (scrawl strip) of Figure 4, produced from the QQ plot in Figure 1, with Figure 1.

(4) Treat the Y data similarly, but interchange the axes and fold along vertical strips.

Steps (1) - (3) produce a strip with a trace in regular jigs (upwards) and jags (downward), as can be seen by inspecting the scrawl plots in Figure 4. Slopes will be preserved, except for a sign change in jags. This means in low absolute slope portions of the trace, there will be shallow appearing "peaks", and, in large absolute slope portions, these "peaks" will appear steeper and denser. Hence information about whether the tails are long or short (compared to the reference distribution) is easily seen on the scrawl plot as a changing density of "peaks", or in more detail as a changing absolute value of the local slope.

An alternative to the folding described in steps (2) and (3) is, at each stage, to directly translate the part of the plot in the remaining uppermost strip down onto the strip below it. This possibility, a *modular* or *sliced strip*, described and illustrated in Veitch (1984), we find less satisfactory.

[2]Source code for the scrawl strip routines (written in the interface language of the statistical package S) is available from the authors on request. (Write to James G. Veitch at Franz, Inc., Suite 270, 1141 Harbor Bay Parkway, Alameda, CA 94501.)

Figure 4 is a scatter plot of auto weight versus price for the same 74 makes of automobile given in Figure 1. It is bordered by two scrawl plots, one for each marginal. The X scrawl plot is presented below the scatter plot, using the same coordinates. The Y scrawl plot is similarly presented alongside the scatter plot. Note that we might instead put the X scrawl plot above the scatter plot and/or the Y scrawl plot on the right side of the scatter plot. We have "ticked" the values of the broadened median, hinges, and eighths and etc., of the marginals in b-letter strips between the scrawl plots and the scatter plot.

The information present in the original QQ plot (Figure 1) is pretty much preserved in the X scrawl strip. Gaps and ties are more obvious; slope changes may be a little less obvious, but one can still see that distribution of auto weights is somewhat short-tailed compared to a Gaussian reference. The bimodality of the weight distribution, with concentrations near 2000 and 3400 pounds, is at least as clear as in the QQ plot. The price distribution is very long-tailed on the high side - - this is apparent even in the scatter plot itself. A tie shows up at the lowest price. Many methods of analysis assume symmetry in the distribution. If one wishes to symmetrize the distribution of prices, given the observed skewness, a commonly suggested procedure is to take logs. The outcome of this procedure is displayed in Figure 5, where one can see from the scrawl strip of log prices that the distribution is *still* somewhat long tailed on the high side. At this point, if one wished to symmetrize or Gaussianize the distribution of prices, it would probably be worth while to work only with the univariate distribution of prices - e.g. one might look at QQ plots or suspended rootograms for different transformations of the price data.

Figure 7 shows scatter plots and scrawl strips of sepal length and sepal width for the famous *Iris* data on which Fisher (1936) illustrated the use of the discriminant function. Only two of the three species are in this plot, *virginica* and *versicolor* (the original data also includes petal length and petal width). We exclude the third species *(setosa)*, as it is clearly differentiated from the other two species on inspection of a scatter plot of all three (see Chambers et al. (1983) page 108, for plots of all three species showing petal width against petal length - *setosa* is easily distinguished, as would be the case here). The large number of ties, due to rounding of the basic measurements for both sepal width and sepal length shows clearly in the scrawl strips. If we wish to avoid distraction of the eye by these ties we may "jitter" each coordinate by adding a small random perturbation to each value (in the example these additions simulate a uniform distribution between 0 and 0.1, so that the resulting intervals abut).

Figure 8 shows shows the scatter plots and scrawl strips for the jittered coordinates. All trace of ties has disappeared and we can see more clearly that sepal length is somewhat skewed to the right. There is also an apparent area of concentration relative to the Gaussian near the lower eighth (5.5 cms) for sepal length, and a gap at about 7.5 cms.

Figure 9, examines the two hard-to-separate *Iris* species, using information from all 4 of the flower dimensions. Each dimension was first put on a relative scale by dividing by the median (for *virginica* and *versicolor* combined) for that coordinate and the results were then assembled. The horizontal scale refers to the natural "size" measure, namely the sum of all four relative dimensions (hence centered near 4), while the vertical scale refers to the natural "shape" measure, the sum of the two relative lengths minus the sum of the two relative widths (hence centered near zero). The horizontal scrawl strip shows smaller peak-to-peak spacings at the ends than in the middle, as would be expected for relative bimodality. The vertical scrawl strip shows a gradation from wide spacing on the - side to narrow spacing on the + side, as would correspond to skewness.

Figure 10 shows the same scatter diagram, with the two species identified. Their overlap in "size" is small - - presumably small enough to account for the appearance of relative bimodality in Figure 9's scrawl strip for size. We can also see that the spread of *versicolor's* shape is much greater than that of *virginica*, and extends much further to negative values than to positive. Even if the distributions of the individual species were Gaussian, this unbalance could account for the sort of skewness apparent in Figure 9's scrawl strip for the shape variate.

This example shows that scrawl strips can be effective in revealing even moderate amounts, either of relative bimodality or of skewness.

Figure 11, using petal length and petal width recentered separately for each of the 3 *Iris* species, offers a synthetic example of the detection of heavy tails due to heterogeneous variability - - both

scrawl plots are consistently steeper in the middle than toward the tails. The reason for the heavy-tailedness is shown in Figure 12, where the recentered results for the 3 species are distinguished. Clearly *versicolor* shows greater variability than the other two (as it should according to Anderson's explanation of its origin by introgressive hybridization of the other two species).

Parameters controlling the appearance of scrawl plots.

In addition to choosing the width of the scrawl plots, one must make two further choices in this procedure. One must choose the reference distribution with which one desires a comparison, possibly because of a wish to visually check on certain assumptions, and one must choose the number of folds, N. The Gaussian distribution will probably be the usual choice, and is the one we have implemented in the plots presented here. For a given set of data, and a given reference distribution, the number of folds, and the overall size of the strip control the visual impact of the scrawl strip. One would like to have scrawl strips of data which come from similar distributions, but with a different sample size, look similar. The number of folds, N, we make does not seem to influence the visual impact very much. A stronger factor often seems to be the visually perceived slopes of the trace in the scrawl strip. To control both N and slope, we would need to make the width of the scrawl strips data dependent; in our basic programs we chose to fix the widths and control the slope.

This choice about visual appearance of the scrawl strip implies that any algorithm to produce a plot must know the physical dimensions of the final scrawl strip. For concreteness, suppose that we wish to produce the X variable scrawl strip, and that this plot will have dimensions w inches in width and h inches in height. Denote the ordered X data by $x_{(1)}, \cdots, x_{(n)}$ and convert these to physical distances from the left edge of the strip (in inches). Denote these distances by $Dx_{(1)}, \cdots, Dx_{(n)}$. Denote the reference scores by $z_{(1)}, \cdots, z_{(n)}$. Denote the index of the lower 25th percentile (the lower quartile) by L and the index of the upper 25th percentile (the upper quartile) by U, so these are given by $x_{(L)}$ and $x_{(U)}$ for the data, $z_{(L)}$ and $z_{(U)}$ for the reference scores respectively. Suppose that the modulus of the visual "slope" of the segments is to be set to approximate s. Our algorithm proceeds in several steps:

(1) Prepare a standard QQ plot with width w, height h' to be determined. We set the distance (in inches) of $z_{(U)}$ to $z_{(L)}$ to be $d = s.(Dx_{(U)}-Dx_{(L)})$, so the visual slope from the lower quartile to the upper quartile is s. Then the total height h' needed is given by

$$h' = d.(z_{(n)}-z_{(1)})/(z_{(U)}-z_{(L)})$$

Note that d depends only on the width of the scrawl strip w and the choice of where to plot the X data, and h' only depends on the shape of the QQ plot, not on its vertical scaling.

(2) Determine the number of strips N needed to fold the plot of step (1) into the physical space allotted, h, by dividing the length, h', found in step (1) by h and rounding up to the nearest integer.

(3) If h' is smaller than h, the QQ plot is merely expanded to height h with no folding. This might happen if the slope constant s were small (usually a poor choice), or if the data distribution were extremely long-tailed (when a different scatter plot is likely to be an improvement).

In examples we have tried, we have found that choosing s=1 or 2 is often reasonable (see Veitch (1984)); the plots given here are in this range (in other circumstances we might prefer to fix the number of folds). The choice of the quartiles (rather than, say, eighths) to determine slope might be argued; however this also seems, from inspection of plots, to work reasonably well.

Some other applications.

Scrawl *strips* - - should not be thought of as restricted to be an appurtenance to scatter plots. Once the tools to make them are available, they can be used anywhere a box plot or other graphical representation or abbreviation of a distribution would be in order. We offer three examples, in the belief that any interested user will be able to recognize and make good use of many more.

All three of these examples involve comparing two or more scrawl strips. For comparisons, it seems desirable to blend the b-letter strip information into the scrawl strip. We shall do this by putting light (dotted or solid) verticals (horizontals if the scrawl is associated with a vertical axis) at the

locations of "b", "e", "m", "e", and "b" (whose unit-Gaussian locations are roughly -1.5, -.85, 0, .85, and +1.5), labelling these verticals unobtrusively, and indicating the other b-letter values by small ticks without labels.

Finally, we could consider two approaches to the use of scrawl strips to show univariate behavior in a "splom" or scatter-plot matrix. In the first, the scrawls are laid in diagonally in the (otherwise empty) boxes on the main diagonal. In the second, each diagonal box shows (two or) three scrawls, laid in horizontally, and referring to that variable and each of those shown in adjacent columns of the matrix. (To make this plot most effective, we may wish (1) to rescale all variables to a common (robust) scale and, possibly, (2) order the variables in a way related to their univariate distribution.

Warning. One potential for trouble with Q-Q plots, as with most exploratory tools, is when an appearance is erroneously regarded as an established truth. This can arise when all three of: (a) it is reasonable to regard the data at hand as a sample, (b) we choose to be concerned with the population, not the sample, *and* (c) we have either not found or not used a significance or confidence procedure (exact *or* approximate) relevant to the appearance that concerns us. (Notice that (a) fails for the "cars" and "same house" examples above.) So far, a useful significance procedure for wigglyness" in a Q-Q plot seems not to be available. (We hope to return to this question, elsewhere.)

Stylized scrawl strips.

For purposes where details like ties and gaps are not needed (are inconsequential and/or confusing), it may pay to stylize one's scrawl strips by (a) using smoothing to fix the corners of the scrawls, (b) plotting each slant of the scrawl as a straight line connecting two vertices, and (c) fixing the number of slants rather than the width of the scrawl strip.

If $< i \mid n >$ represents the reference position for the ith of a sample of n, let m^* be the least integer as large as $1.1\sqrt{n}$, and plan to have m^* slants by taking the height (before folding) of a slant to be $D = (< n \mid n > - < 1 \mid n >)/m^*$. We then find the internal vertices by regressing x_i on $< i \mid n >$ for i's satisfying

$$< 1 \mid n > + (j - \frac{1}{2})D \leq < i \mid n > \leq < 1 \mid n > + (j + \frac{1}{2})D$$

and inserting $< 1, n > + jD$ in the regression to find the jth vertex, $j = 1, 2, \ldots, m^{-1}$. These vertices are then plotted alternately at the top and bottom of the scrawl, and connected by straight lines. The points falling in the end slants are then plotted individually as before.

Conclusions.

Presenting marginal data distributions on the edges of a bivariate scatter plot by providing b-letter strips (or letter strips) to mark the b-letter values (or letter values), allows considerable information about symmetry, center and spread of these marginals to be presented in minimal space. Such letter (or b-letter) strips may be used to advantage in other types of plots as well (e.g. in QQ plots). Additional use of scrawl strips on the edges of a scatter plot allows other features of the marginal data distributions to be easily seen together with the original data. Since scrawl strips are modified QQ plots they inherit useful features of QQ plots, including information about the shape of the marginal distribution, the presence and type of skewness, the presence of heavy tails, gaps and ties in the ordered data, appearances of bimodality, and relation to Gaussianity. This type of information is either absent or depicted poorly both in the original scatter plot and also in most univariate presentations such as histograms, rootograms, boxplots, and simply "ticking" in or jittering in the marginal values.

Acknowledgements

This work was initially supported by U.S. Department of Energy contract DEAC0281ER10841 with Princeton University. This contract and a National Science Foundation grant MCS8304372 supported the Computing equipment used at Princeton. J. W. Tukey's later work was supported by the Army Research Office (Durham) through contracts DAAG29-82-K-0178, DAAL03-86-K-0073, and DAAL03-88-K-0045.

References

Chambers, J., Cleveland, W., Kleiner, B., Tukey, P. A., 1983, *Graphical Methods for Data Analysis*, Duxbury Press, Boston.

Fisher, R. A., 1936, Ann. Eugen., Vol 7, pp. 179-188.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (editors), 1983b, *Understanding Robust and Exploratory Data Analysis*, John Wiley, New York.

Hoaglin, D.C., Mosteller, F., and Tukey, J. W. (editors), 1985b, *Exploring Data Tables, Trends and Shapes*, John Wiley, New York.

Mendoza, C., 1984, *Simple approximate variances for bletter values; a sampling enquiry* Tech. Report 269, Series 2, Dept. of Stat., Princeton University (includes an Appendix by J. W. Tukey).

Tufte, E., 1983, *The Visual Display of Quantitative Information*, Graphics Press, Cheshire, Connecticut.

Tukey, J.W., 1965l, "The future of processes of data analysis, *Proceedings of the Tenth Conference on the Design of Experiments in Army* Research Development and Testing, (ARO-D Report 65-3), pp. 691-729.

Tukey, J.W., 1977a, *Exploratory Data Analysis*, Addison Wesley.

Veitch, J., 1984, *Scrawl Plots and Letter and Bletter strips*, Tech. Report 268, Series 2, Dept. of Stat., Princeton University.

Appendix: Extending a finite batch to an unlimited distribution.

If $y_n \le y_{n-1} \le , \ldots , \le y_2 \le y_1$, is an ordered batch, and there are not ties, it is convenient

(1) to assign $p_i = (3i\text{-}1)/(3n+1)$ to y_i, and

2) to interpolate linearly between (p_i , y_i) and (p_{i+1}, y_{i+1}).

It remains then to deal with p-values outside $[y_1, y_n]$ and (b) with ties.

A reasonable choice is to take m as the integer part of $\sqrt{n}$, and to extrapolate (p_1, y_1) and (p_m , y_m) *exponentially* for smaller (p, y). This means taking

(*) $\quad \ln p = A + By$

fitting A and B to (p_1, y_1) and (p_m , y_m) and then using (*) for $p < p_1, y < y_1$.

A similar extrapolation for (p_n , y_n) (p_{n+1-m} , y_{n+1-m}) applies for $p > p_n , y > y_n$.

For ties, suppose that $y_j = y_{j+1}=...=y_k$ is the full extent of one tie. Let $(p_m \cdot y_m)$ correspond to $((j+k)/2)$ the median of the tie. Put

$$y_L = (y_{j-1} + y_j)/2), y_R = (y_R + y_{R+1})/2 ,$$

$$p_L = (p_{j-1}-p_j), p_R = (p_R+p_{R+1})/2$$

and plan to do linear interpolation between (p_L , y_L) and (p_m , y_m) and al~o between (p_m , y_m) and (p_R , y_R). (If the tying is due to a regular pattern of possible values we should replace y_{j-1} by the next lower possible value and y_{k+1} by the next higher possible value.)

If there are no ties, $(p_m , y_m) = (p_j , y_j)$ and (p_L , y_L) is the midpoint of the segment joining (p_{j-1}, y_{j-1}) to (p_j , y_j) so that interpolation between (p_L , y_L) and (p_M , y_m) is, in this special case the same as that between (p_{j-1}, y_{j-1}) and (p_j , y_j). Thus the procedure for ties reduces gracefully to the procedure for no ties - - and can, if it is more convenient, be always applied.

For the small-group large-group case, where (q_n , z_n) may represent the smaller-group's observations with

$$q_n = \frac{3h-1}{3n_z+1} \quad \text{but} \quad p_i = \frac{3i-1}{3n_y+1}$$

NOTE: Letters used with years on John Tukey's publications correspond to bibliographies in all volumes of his collected papers.

where $n_z < n_y$, we find a standard value for z_h by equating q_h and p_i, giving

$$i = \frac{1}{3} + \frac{1}{3}(3n_y+1)p_i = \frac{1}{3} + \frac{1}{3}(3n_y+1)\frac{3h-1}{3n_z+1}$$

$$= \frac{1}{3} + \frac{3n_y+1}{3n_z+1}(h-\frac{1}{3})$$

and then interpolating for i. (We may have to face ties; we will not have to extrapolate.)

In our example, California cities with ranks (from above) 22 to 25 had 51.5% "stayers". Further $n_z = 73$, $n_y = 475$. The median rank of 23.5 converts to

$$i = 1 + \frac{1426}{220}(23.5 - \frac{1}{3}) = 151.16$$

In the non-California distribution 151 falls on a 53.7% tie, while 152 falls a 53.6% tie. Thus (151.5, 53.65%) and (150, 53.70%) are to be interpolated to 151.16, giving 53.66% as the standard value to be used in the single-scrawl Figure 17.

Figure 1

QQ plot for weights of 74 autos in the 1979 model year

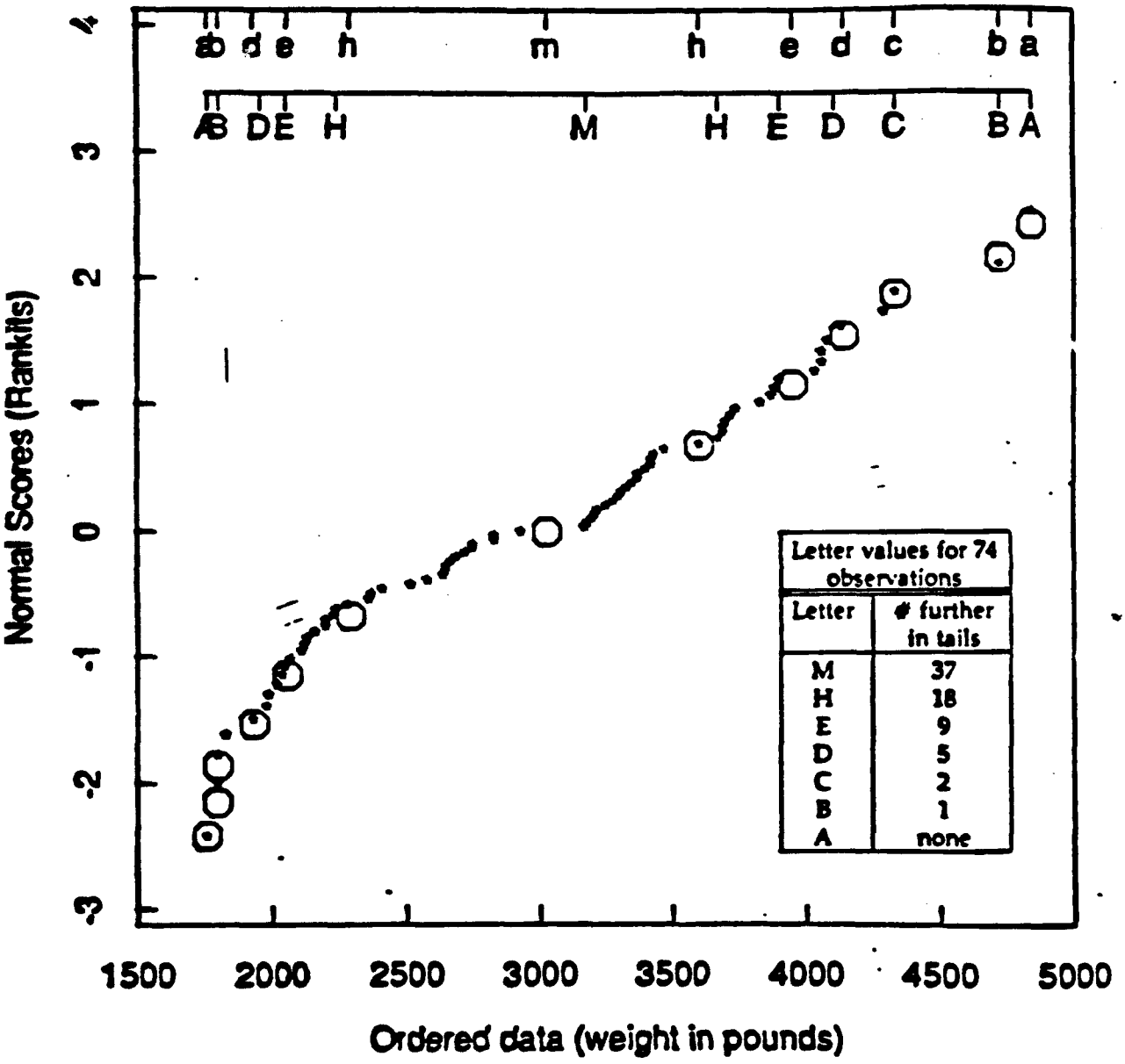


Figure 2

Histogram of weights of 74 autos
in the 1979 model year

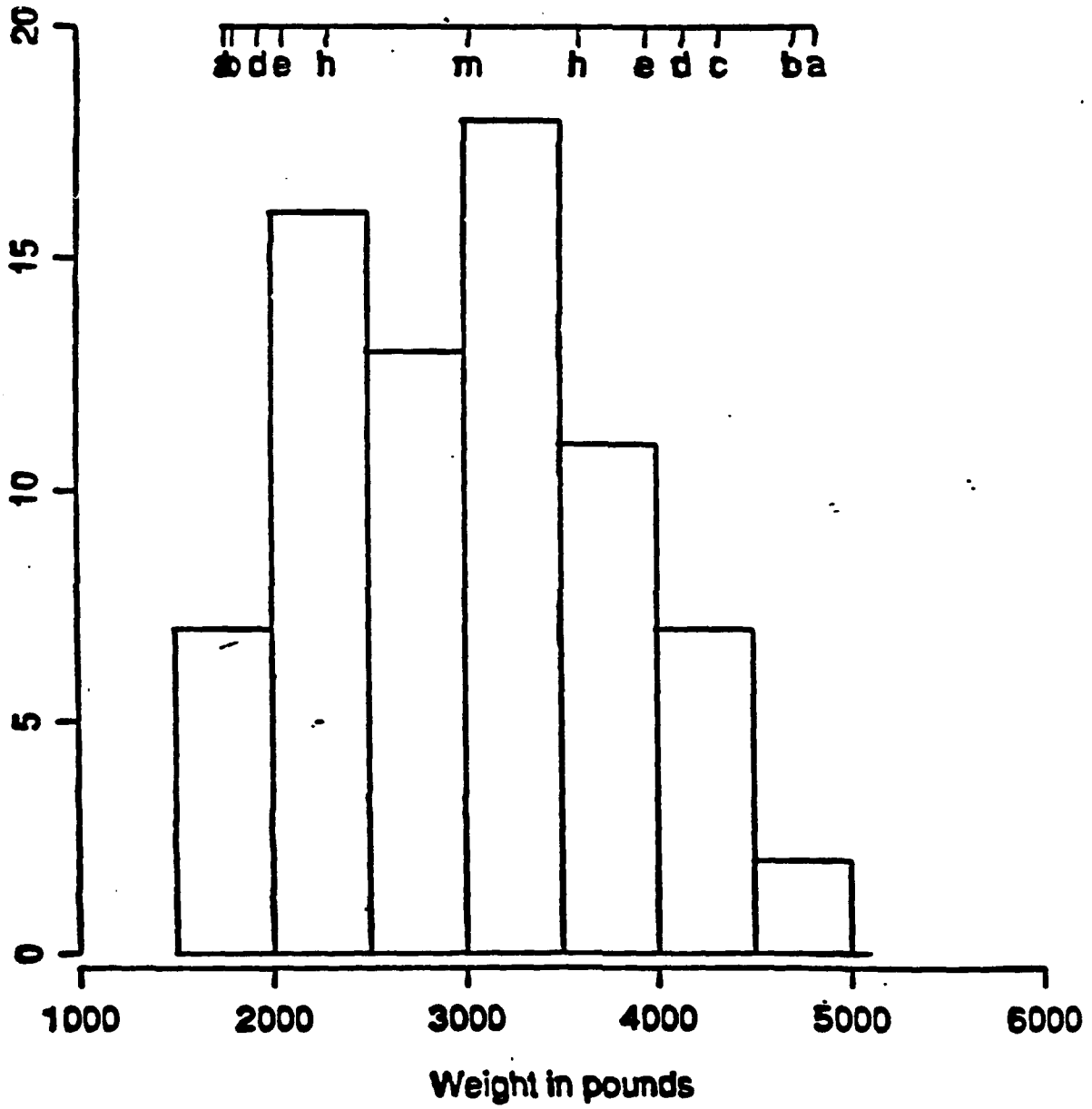


Figure 3

QQ plot after the first fold

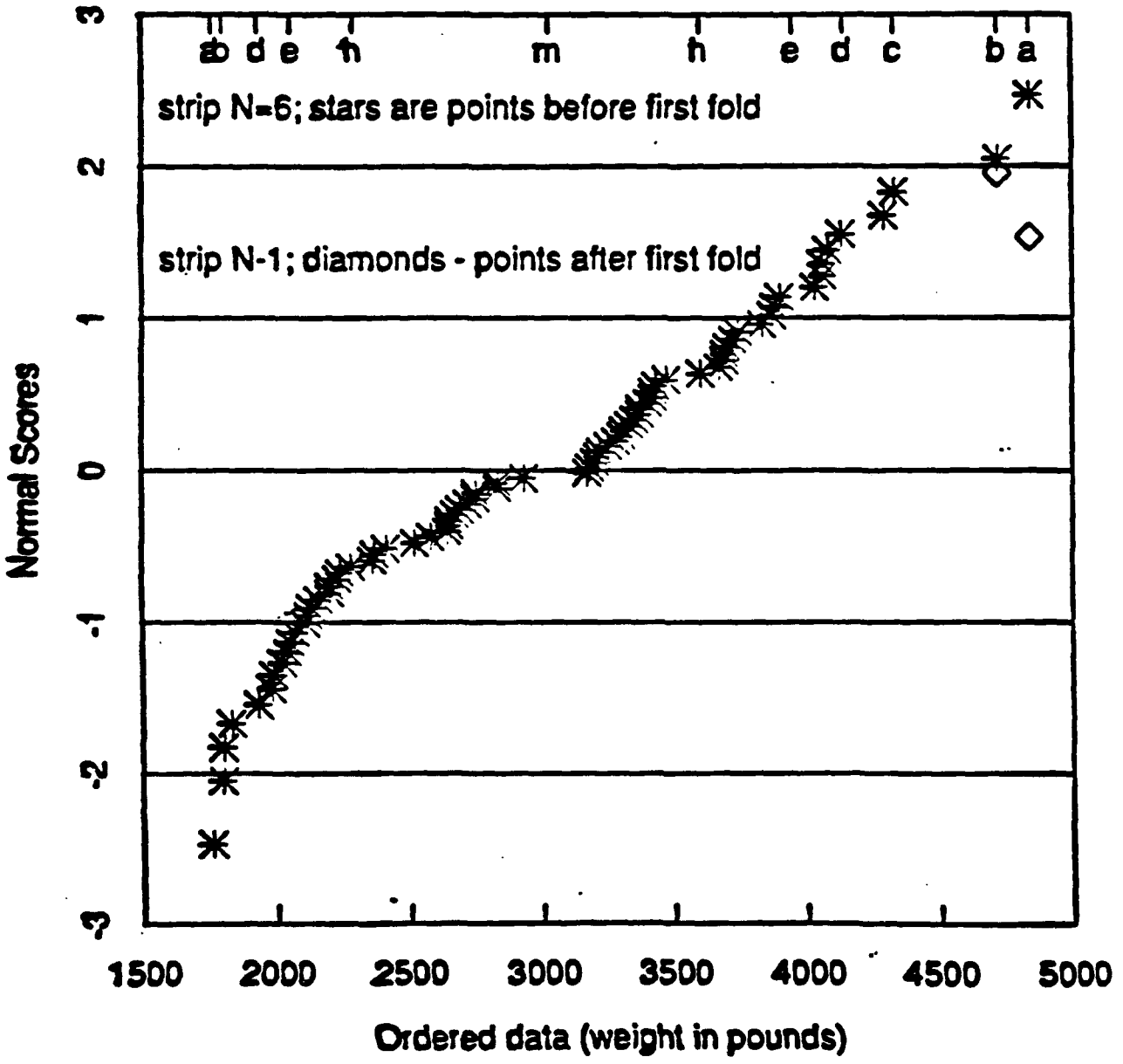


Figure 4

Statistics for 74 auto models of 1979

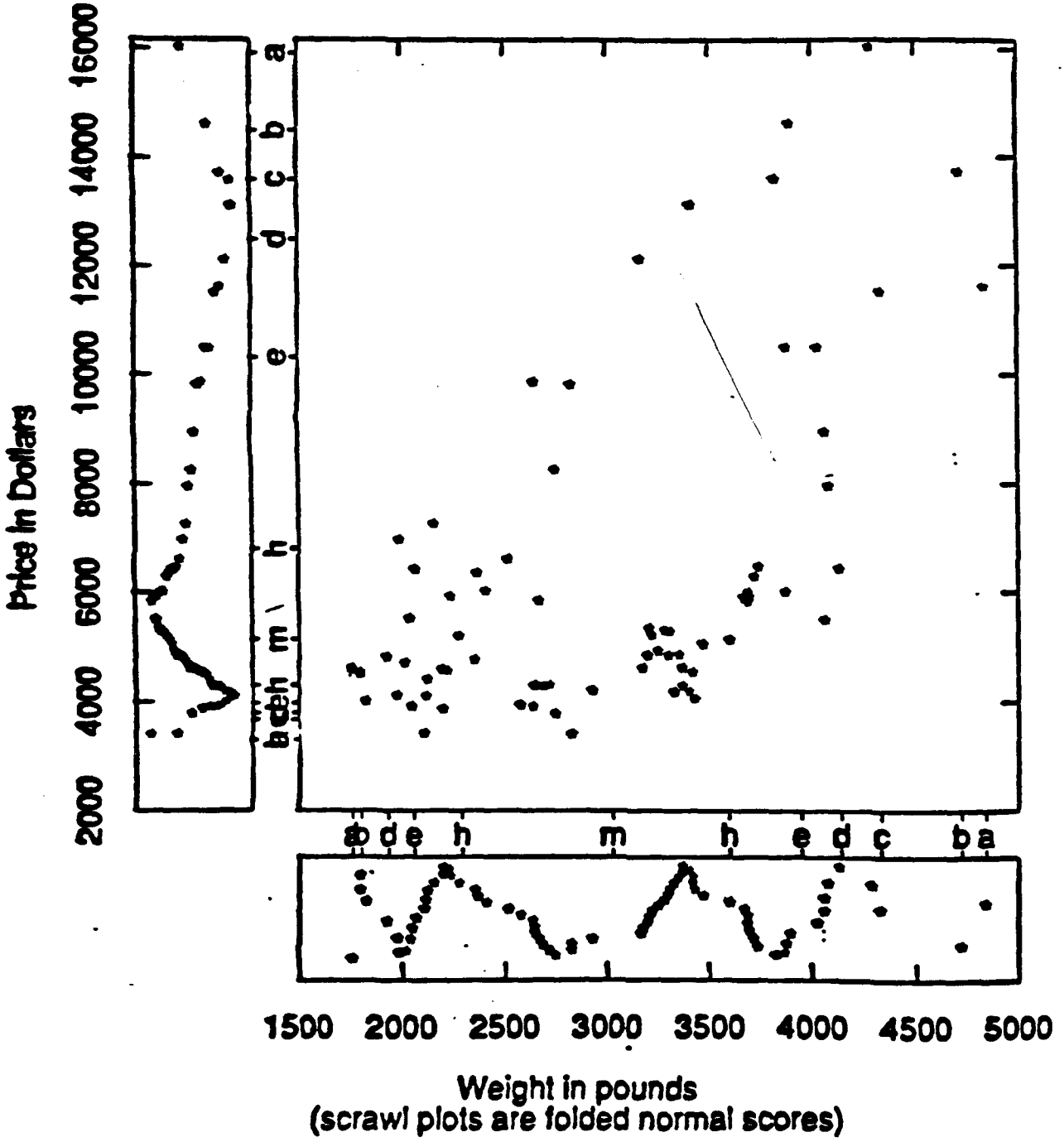


Figure 5

Statistics for 74 auto models of 1979

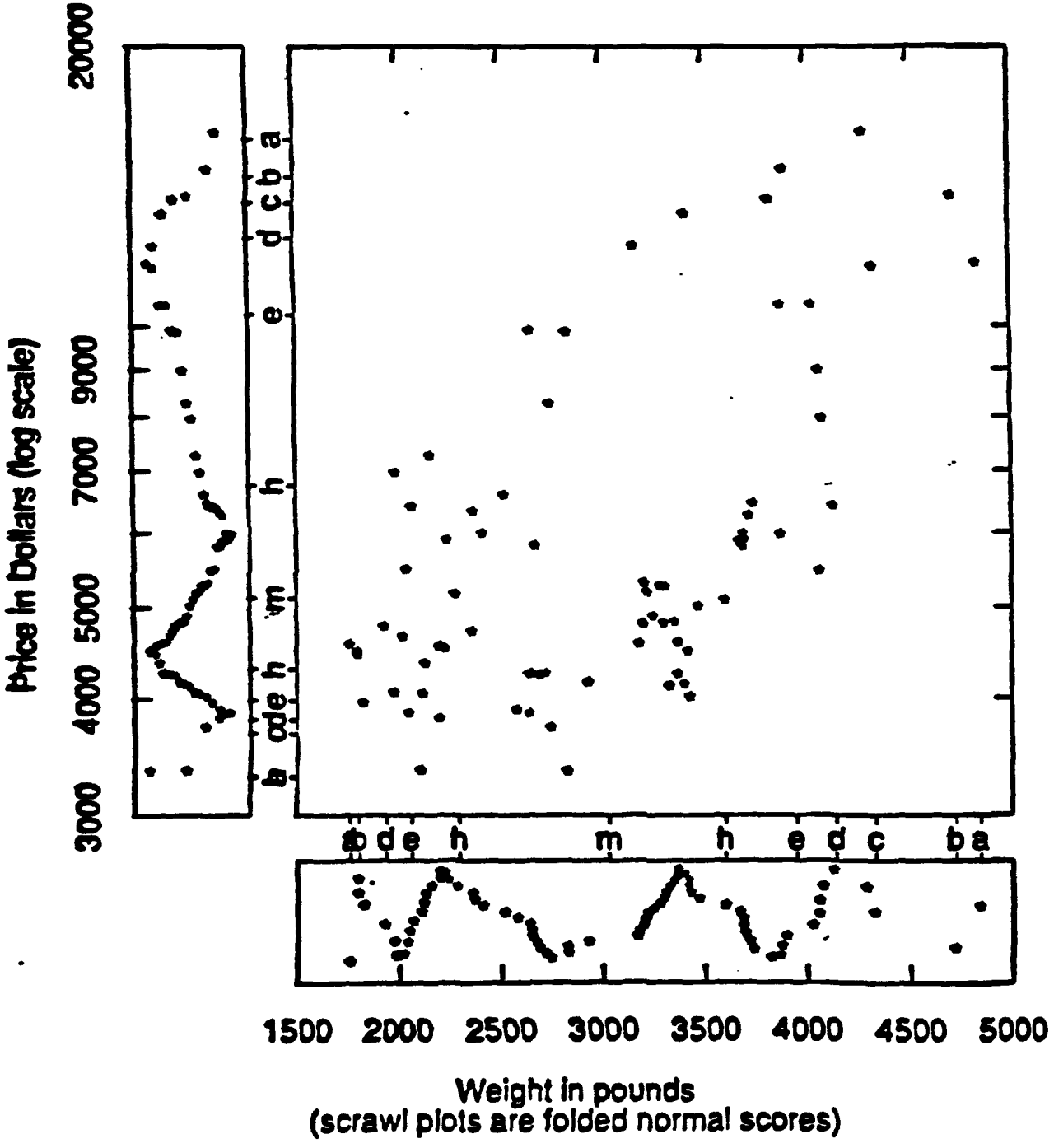


Figure 6

QQ plot for weights of 74 autos
in the 1979 model year

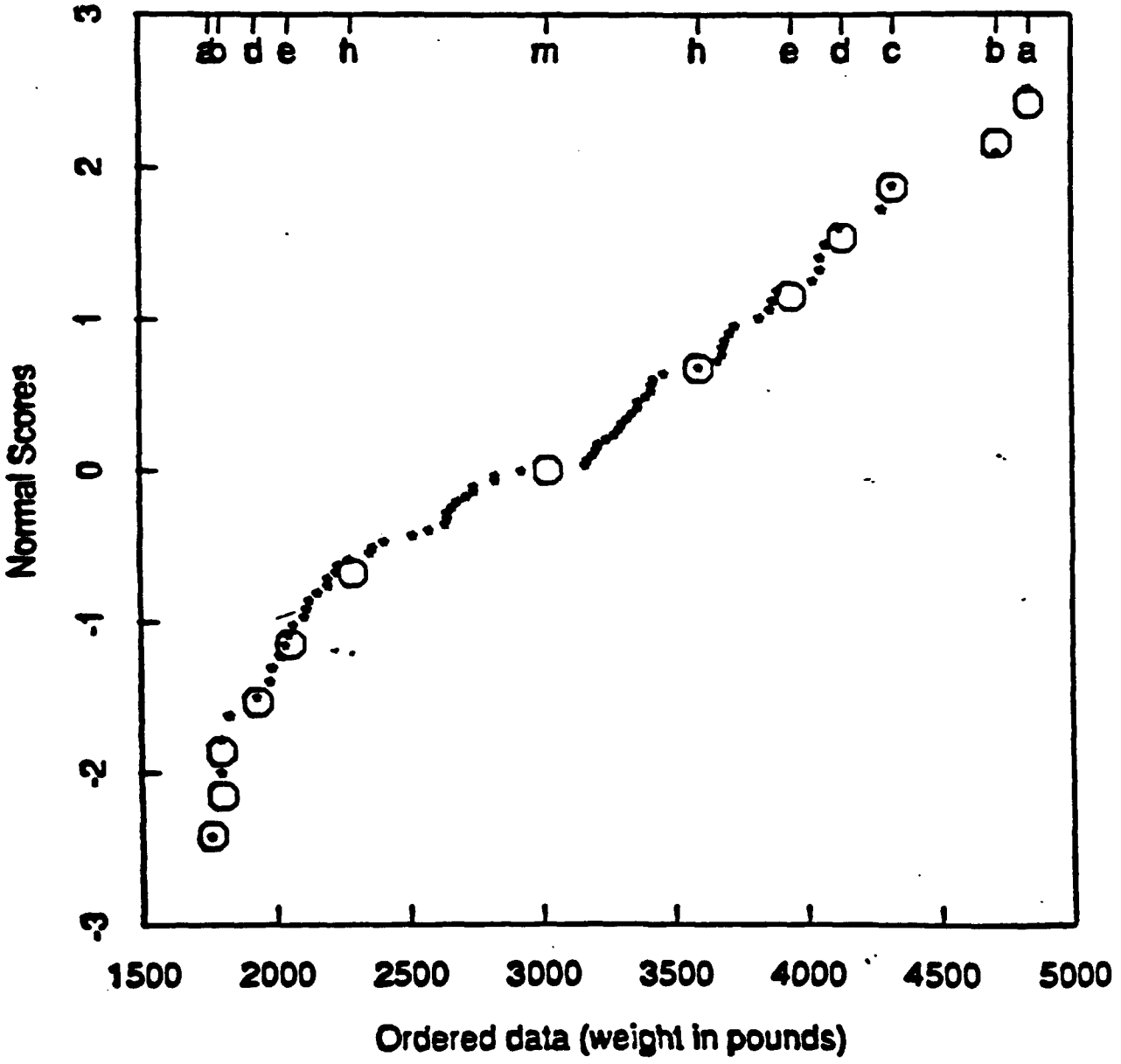


Figure 7

Fisher's Iris data
(virginica and versicolor)

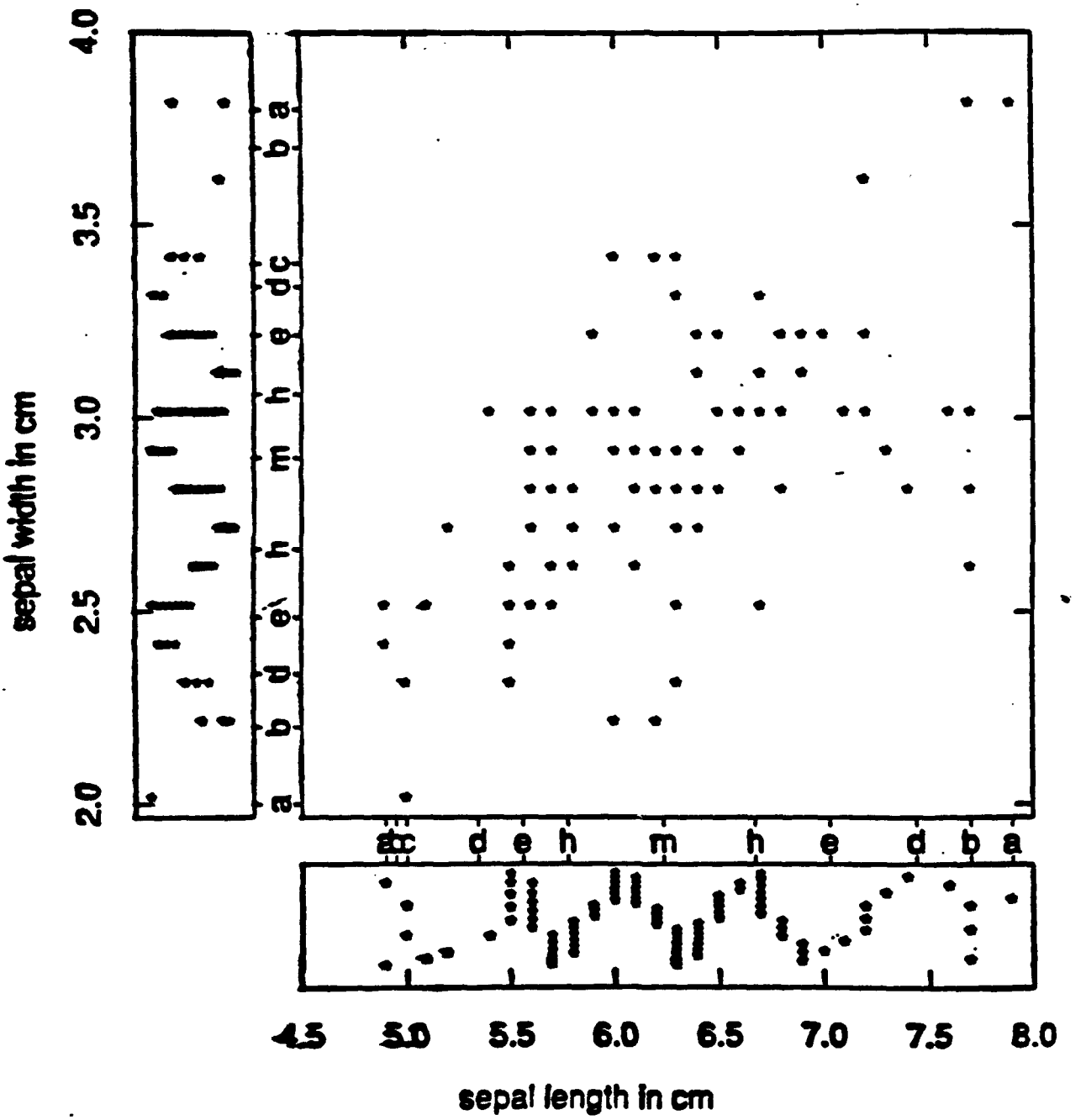


Figure 8

Fisher's Iris data (jittered)

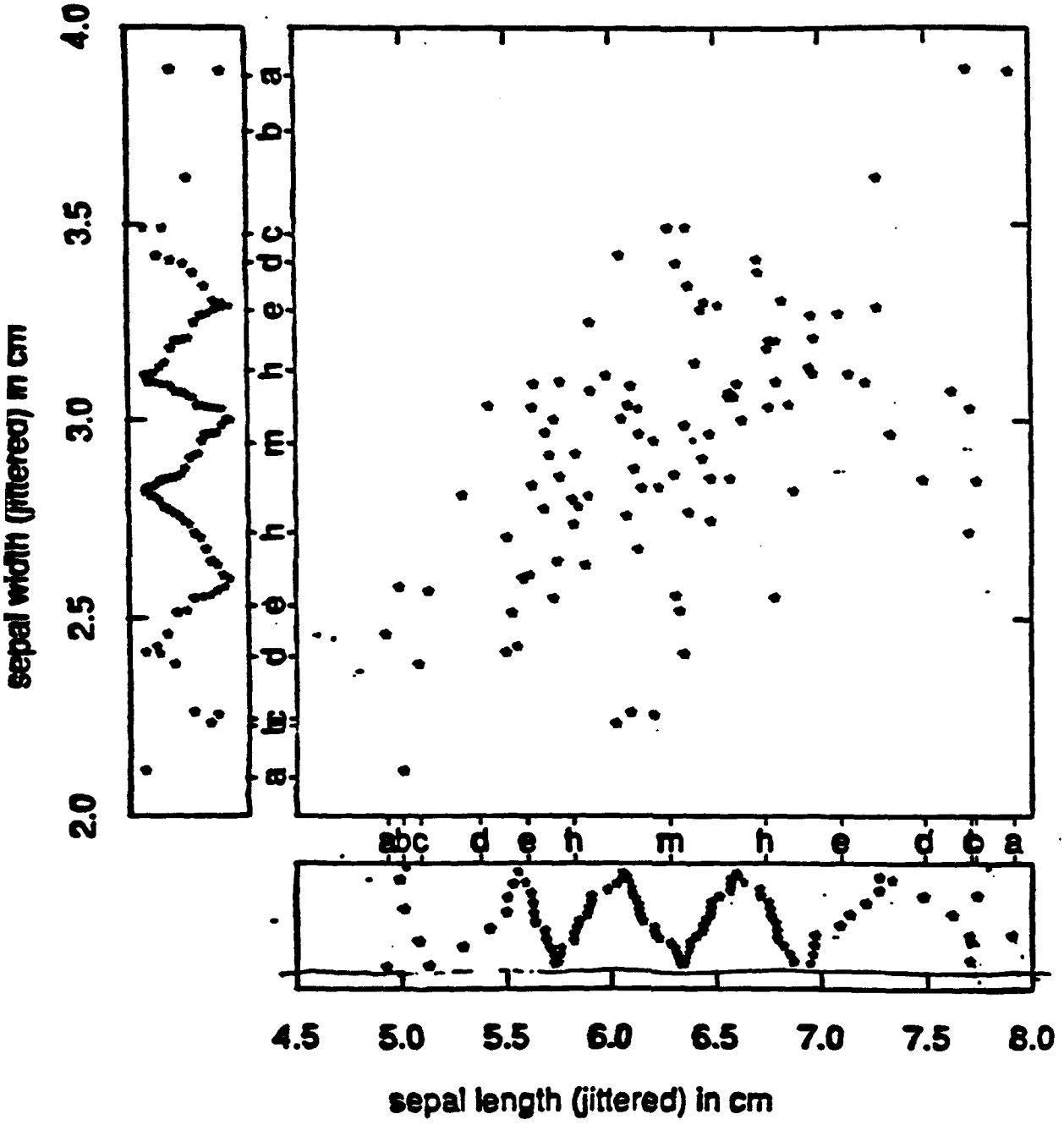


Figure 9

Transformed Iris data
(virginica and versicolor)

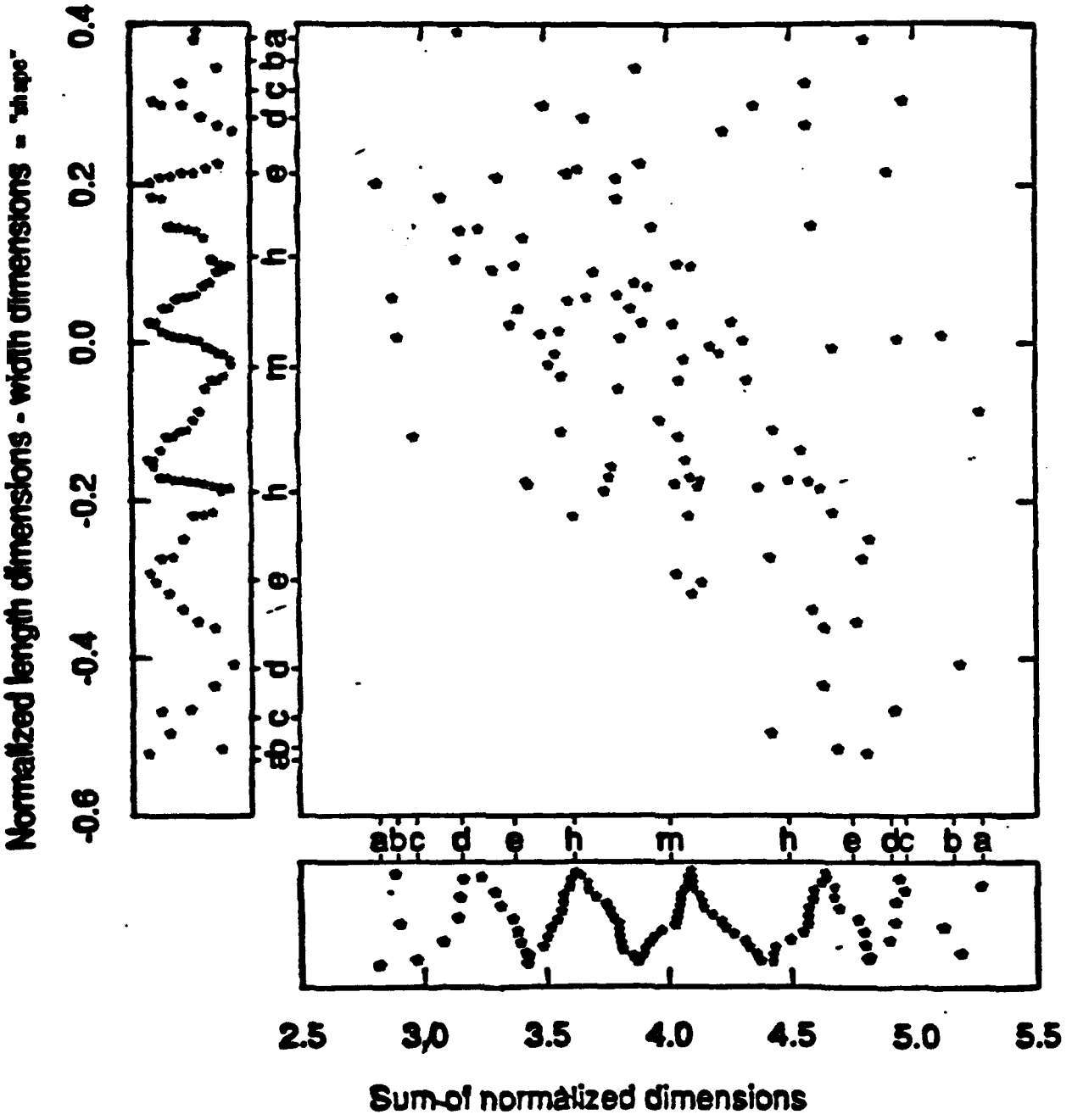


Figure 10

Transformed iris data
(Stars plot virginica, diamonds plot versicolor)

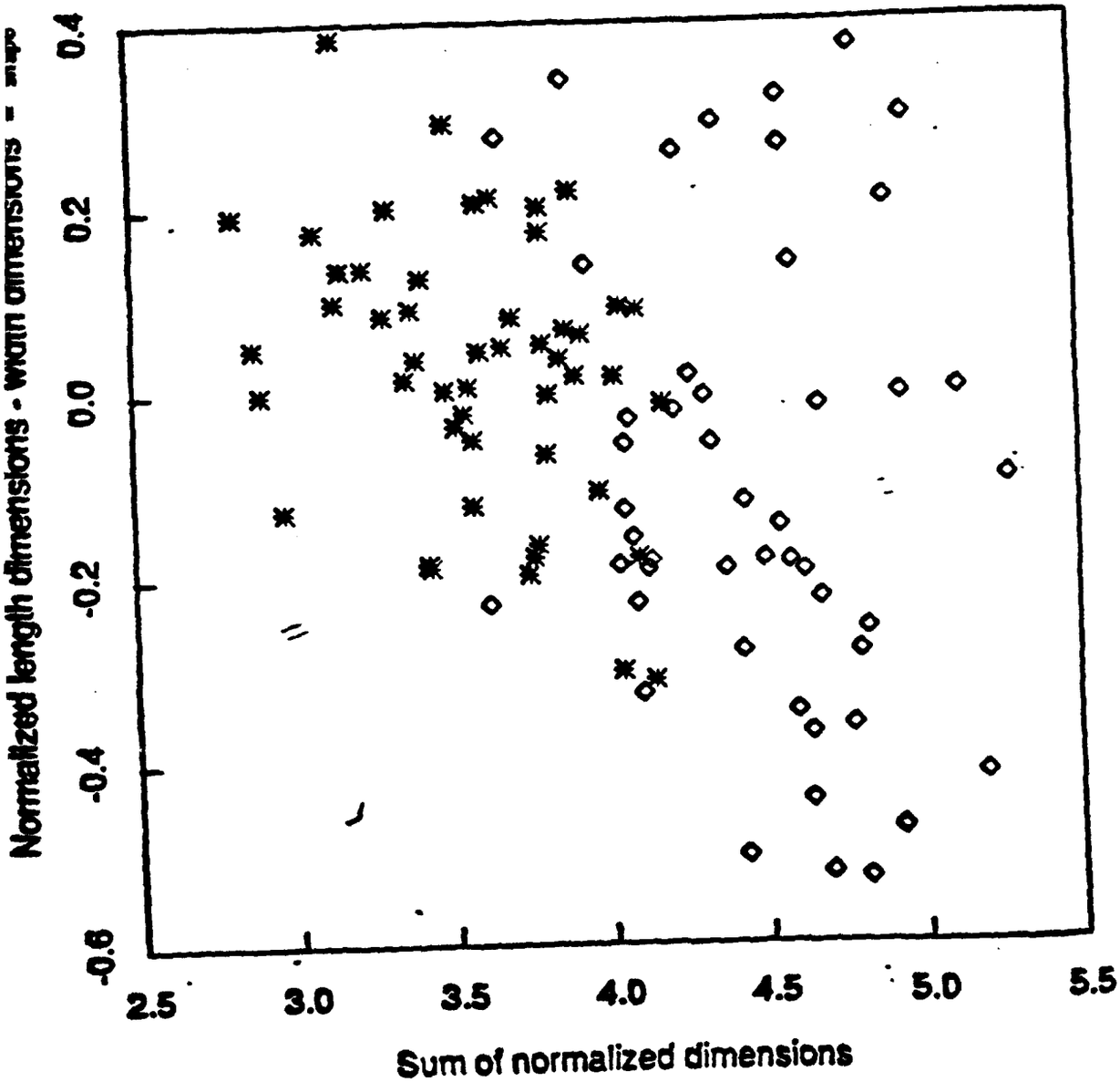


Figure 11

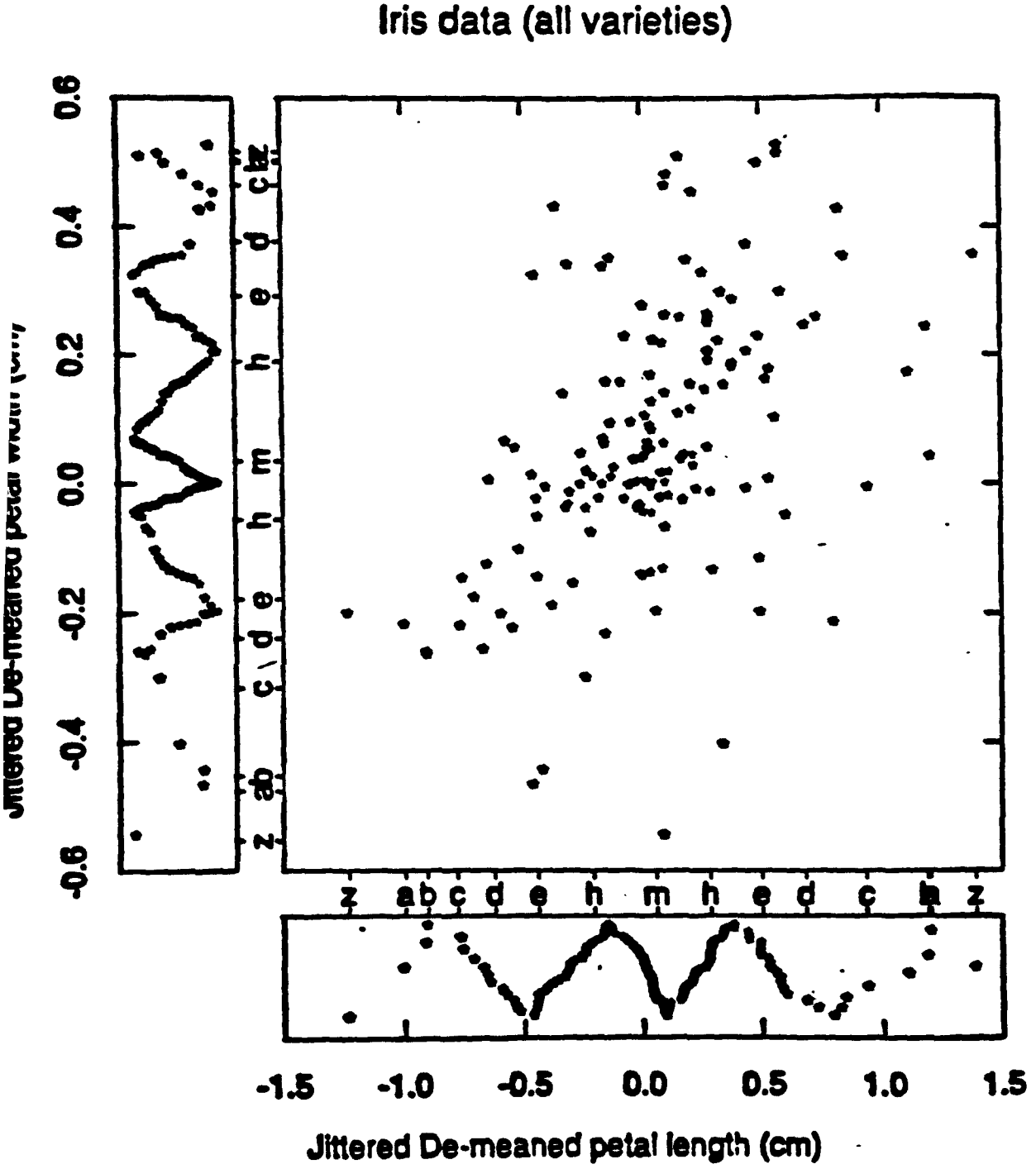
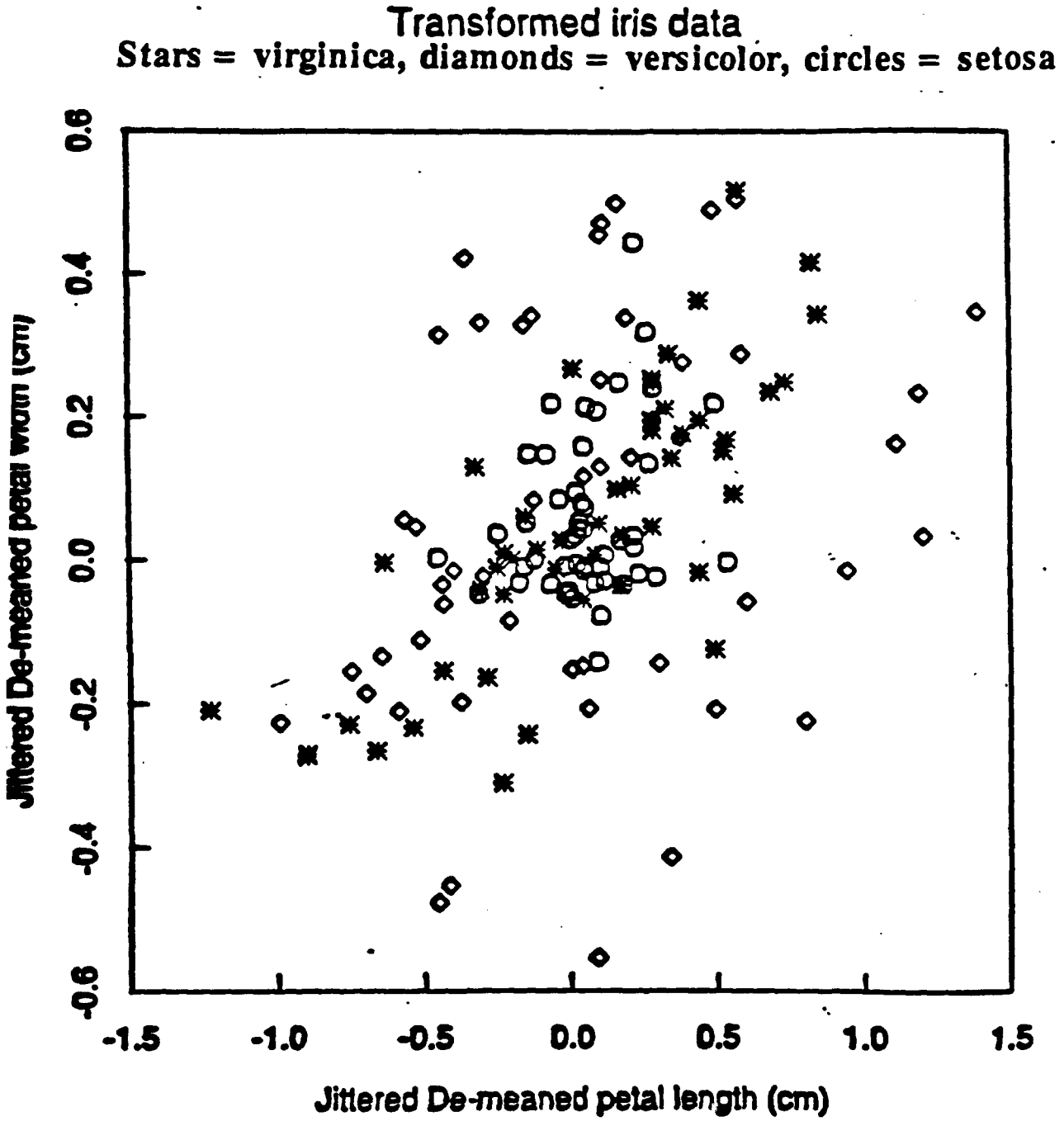

Iris data (all varieties)
Jittered De-meaned petal width (cm)
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
z ab c d e h m h e d c b z
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Jittered De-meaned petal length (cm)

Figure 12


Transformed iris data
Stars = virginica, diamonds = versicolor, circles = setosa
Jittered De-meaned petal width (cm)
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Jittered De-meaned petal length (cm)